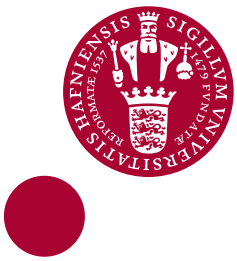

Learning stable and predictive structures in kinetic systems

Pfister, Niklas; Bauer, Stefan; Peters, Jonas

Published in:
Proceedings of the National Academy of Sciences of the United States of America

DOI:
10.1073/pnas.1905688116

Publication date:
2019

Document version
Peer reviewed version

Learning stable and predictive structures in kinetic systems

Niklas Pfister[a,1], Stefan Bauer[b], and Jonas Peters[c]

[a]Seminar for Statistics, Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland; [b]Empirical Inference, Max-Planck-Institute for Intelligent Systems, 72076 Tübingen, Germany; and [c]Department of Mathematical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark

Edited by Bin Yu, University of California, Berkeley, CA, and approved November 1, 2019 (received for review April 4, 2019)

Learning kinetic systems from data is one of the core challenges in many fields. Identifying stable models is essential for the generalization capabilities of data-driven inference. We introduce a computationally efficient framework, called CausalKinetiX, that identifies structure from discrete time, noisy observations, generated from heterogeneous experiments. The algorithm assumes the existence of an underlying, invariant kinetic model, a key criterion for reproducible research. Results on both simulated and real-world examples suggest that learning the structure of kinetic systems benefits from a causal perspective. The identified variables and models allow for a concise description of the dynamics across multiple experimental settings and can be used for prediction in unseen experiments. We observe significant improvements compared to well-established approaches focusing solely on predictive performance, especially for out-of-sample generalization.

kinetic systems | causal inference | stability | invariance | structure learning

Quantitative models of kinetic systems have become a cornerstone of the modern natural sciences and are universally used in scientific fields as diverse as physics, neuroscience, genetics, bioprocessing, robotics, or economics (1–5). In systems biology, mechanistic models based on differential equations, although not yet standard, are being increasingly used, for example, as biomarkers to predict patient outcomes (6), to improve predicting ligand-dependent tumors (7), or for developing mechanism-based cancer therapeutics (8). While the advantages of a mechanistic modeling approach are by now well established, deriving such models by hand is a difficult and labor-intensive manual effort. With new data acquisition technologies (9–11), learning kinetic systems from data has become a core challenge.

Existing data-driven approaches infer the parameters of ordinary differential equations by considering the goodness of fit of the integrated system as a loss function (12, 13). To infer the structure of such models, standard model selection techniques and sparsity-enforcing regularizations can be used. When evaluating the loss function or performing an optimization step, these methods rely on numerically integrating the kinetic system. There are various versions, and here we concentrate on the highly optimized Matlab implementation Data2Dynamics (14). It can be considered as a state-of-the-art implementation for directly performing an integration in each evaluation of the loss function. However, even with highly optimized integration procedures, the computational cost of existing methods is high and depending on the model class, these procedures can be infeasible. Moreover, existing data-driven approaches, not only those using numerical integration, infer the structure of ordinary differential equations from a single environment, possibly containing data pooled from several experiments, and focus solely on predictive performance. Such predictive-based procedures have difficulties in capturing the underlying causal mechanism and, as a result, they may not predict well the outcome of experiments that are different from the ones used for fitting the model.

Here, we propose an approach to model the dynamics of a single target variable rather than the full system. The resulting computational gain allows our method to scale to systems with many variables. By efficiently optimizing a noninvariance score our algorithm consistently identifies causal kinetic models that are invariant across heterogeneous experiments. In situations where there is not sufficient heterogeneity to guarantee identification of a single causal model, the proposed variable ranking may still be used to generate causal hypotheses and candidates suitable for further investigation. We demonstrate that our framework is robust against model misspecification and the existence of hidden variables. The proposed algorithm is implemented and available as an open-source R package. The results on both simulated and real-world examples suggest that learning the structure of kinetic systems benefits from taking into account invariance, rather than focusing solely on predictive performance. This finding aligns well with a recent debate in data science proposing to move away from predictability as the sole principle of inference (15–21).

Results

Predictive Models versus Causal Models. Established methods mostly focus on predictability when inferring biological structure from data by selecting models. This learning principle, however, does not necessarily yield models that generalize well to unseen experiments, since purely predictive models remain

Significance

Many real-world systems can be described by a set of differential equations. Knowing these equations allows researchers to predict the system's behavior under interventions, such as manipulations of initial or environmental conditions. For many complex systems, the differential equations are unknown. Deriving them by hand is infeasible for large systems, and data science is used to learn them from observational data. Existing techniques yield models that predict the observational data well, but fail to explain the effect of interventions. We propose an approach, CausalKinetiX, that explicitly takes into account stability across different experiments. This allows us to draw a more realistic picture of the system's underlying causal structure and is a first step toward increasing reproducibility.

Author contributions: N.P., S.B., and J.P. designed research; N.P., S.B., and J.P. performed research; N.P., S.B., and J.P. contributed new reagents/analytic tools; N.P. analyzed data; and N.P., S.B., and J.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Well-documented R and python code and data are available at http://CausalKinetiX.org. Code is also available as an open-source R package on CRAN (http://cran.r-project.org/web/packages/CausalKinetiX).

[1]To whom correspondence may be addressed. Email: niklas.pfister@stat.math.ethz.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1905688116/-/DCSupplemental.

First published November 27, 2019.

agnostic with respect to changing environments or experimental settings. Causal models (22, 23) explicitly model such changes by the concept of interventions. The principle of autonomy or modularity of a system (24, 25) states that the mechanisms which are not intervened on remain invariant (or stable). This is why causal models are expected to work more reliably when predicting under distributional shifts (26–28).

Causality through Stability. In most practical applications the causal structure is unknown, but it may still be possible to infer the direct causes of a target variable Y, say, if the system is observed under different, possibly unspecified experimental settings. For nondynamical data, this can be achieved by searching for models that are stable across all experimental conditions; i.e., the parameter estimates are similar. Covariates that are contained in all stable models, i.e., in their intersection, can be proved to be causal predictors for Y (17, 29). The intersection of stable models, however, is not necessarily a good predictive model. In this work, we propose a method for dynamical systems that combines stability with predictability, we show that the inferred models generalize to unseen experiments, and we formalize its relation to causality (*Materials and Methods*).

CausalKinetiX: Combining Stability and Predictability. The observed data consist of a target variable Y and covariates X measured at several time points across different experimental setups and are assumed to be corrupted with observational noise (Fig. 1, *Top*). Our proposed method, CausalKinetiX, exploits the assumption that the model governing the dynamics of Y remains invariant over the different experiments. We assume there is a subset S^* of covariates, such that for all n repetitions, and $\frac{d}{dt}Y_t$ depends on the covariates in the same way; i.e.,

$$\frac{d}{dt}Y_t^{(i)} = f\left(X_t^{S^*,(i)}\right), \text{ for all } i = 1, \ldots, n. \qquad [1]$$

The covariates are allowed to change arbitrarily across different repetitions i. Instead of fitting based only on predictive power, CausalKinetiX explicitly measures and takes into account violations of the invariance in **[1]**. Fig. 1 depicts the method's full workflow. It ranks a collection of candidate models $\mathcal{M} = \{M_1, \ldots, M_m\}$ for the target variable (*Materials and Methods*) based both on their predictive performance and whether the invariance in **[1]** is satisfied. For a single model, e.g., $\frac{d}{dt}Y_t^{(i)} = \theta X_t^{8,(i)}$, and noisy realizations $\widetilde{Y}_{t_1}^{(i)}, \ldots, \widetilde{Y}_{t_L}^{(i)}$, we propose to compare the 2 data fits illustrated in Fig. 2. Data fit A calculates a smoothing spline to the data using all realizations from the same experiment; see **[3]**. This fit serves only as a baseline for comparison: It does not incorporate the form of the underlying kinetic model, but is entirely data driven. To obtain data fit B, we fit the considered model, $\frac{d}{dt}Y_t = \theta X_t^8$, on the data from all other experiments (explicitly leaving out the current experiment) by regressing estimated derivatives on the predictor variables. In this example, the model is linear in its parameters,

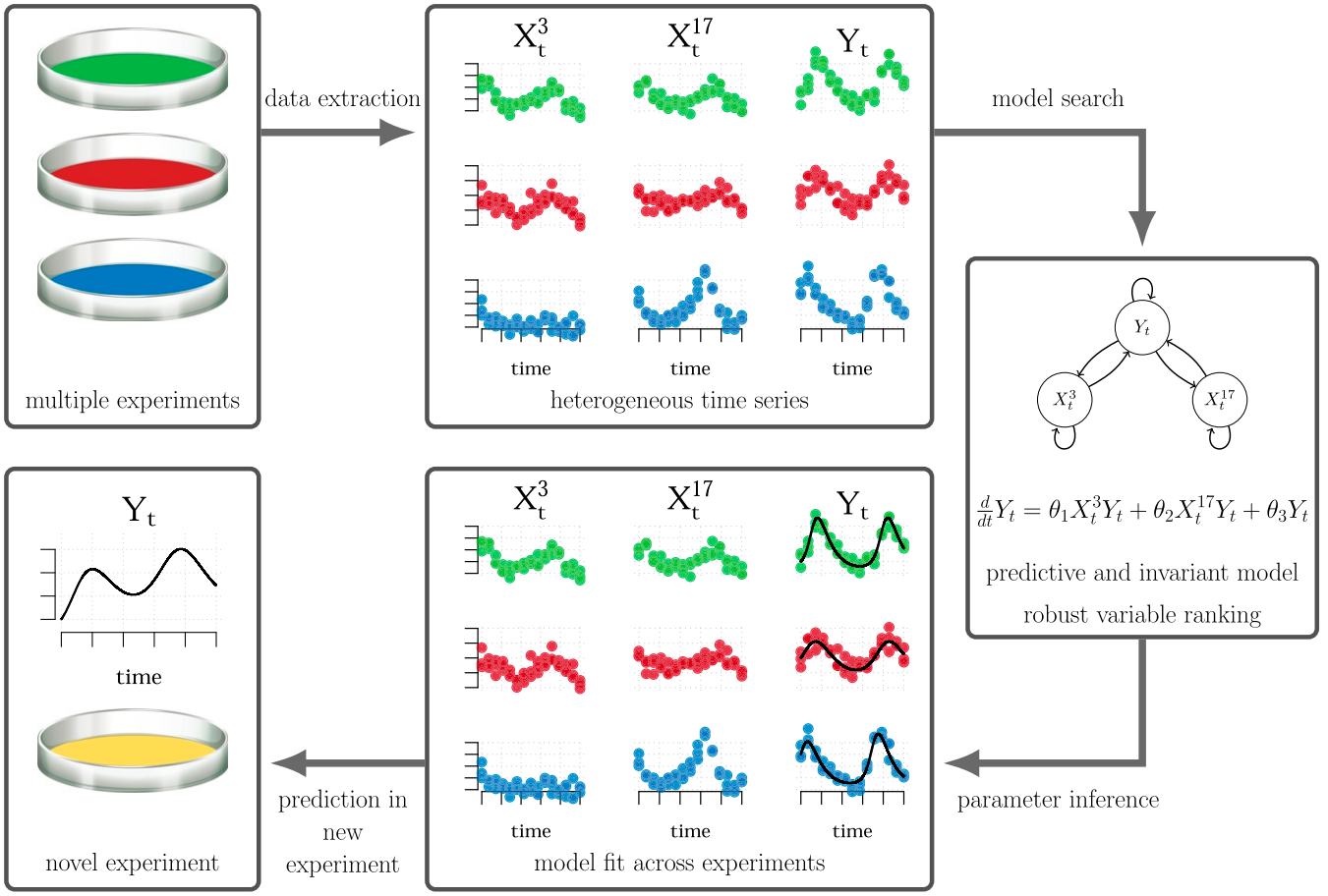

Fig. 1. The framework of CausalKinetiX. The data for target variables *Y* and predictors *X* come from different experiments; we rank models according to their ability to fit the target well in all experiments; the top-ranked model is then fitted to the data; it allows us to predict the target in an unseen experiment.

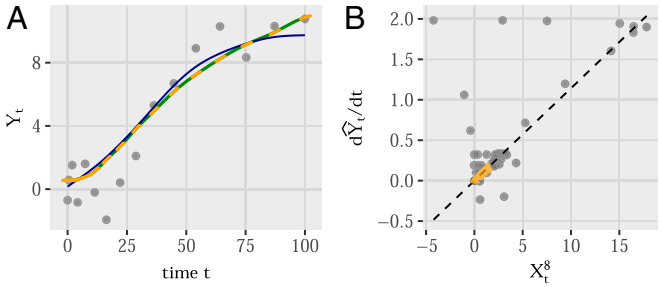

Fig. 2. CausalKinetiX assigns a score to each model that trades off predictability and invariance. Here, we consider the model $dY_t/dt = \theta X_t^8$. (*A*) For each realization, 2 data fits are considered. An entirely data-driven nonlinear smoother (data fit A, blue) is compared against a model-based fit (data fit B, green) with constraints on the derivatives (orange lines). (*B*) The derivative constraints are obtained from all other experiments: They correspond to fitted values (orange triangles) in a regression of the estimated derivatives on the predictors.

and it therefore suffices to use linear regression. Data fit B fits a smoothing spline to the same data, subject to the constraint that its derivatives coincide with the fitted values from the regression inferred solely based on the other experiments; see **[5]**. Data fits A and B are compared by considering the goodness of fit for each realization $i = 1, \dots, n$. More specifically, each model $M \in \mathcal{M}$ obtains, similar in spirit to ref. 30, the noninvariance score

$$T(M) := \frac{1}{n} \sum_{i=1}^{n} \left[\mathrm{RSS}_B^{(i)} - \mathrm{RSS}_A^{(i)} \right] / \left[\mathrm{RSS}_A^{(i)} \right],$$

where $\mathrm{RSS}_*^{(i)} := \frac{1}{L} \sum_{\ell=1}^{L} (\hat{y}_*^{(i)}(t_\ell) - \widetilde{Y}_{t_\ell}^{(i)})^2$ is the residual sum of squares based on the respective data fits $\hat{y}_A^{(i)}$ and $\hat{y}_B^{(i)}$. Due to the additional constraints, RSS_B is always larger than RSS_A.

The score is large either if the considered model does not fit the data well or if the model's coefficients differ between the experiments. Models with a small score are predictive and invariant. These are models that can be expected to perform well in novel, previously unseen experiments. Models that receive a small residual sum of squares, e.g., because they overfit, do not necessarily have a small score T. We will see in *Generalization in Metabolic Networks* that such models may not generalize as well to unseen experiments. This assumes that **[1]** holds (approximately) when including the unseen experiments, too. Naturally, if the unseen experiments may differ arbitrarily from the training experiments, neither CausalKinetiX nor any other method will be able to generalize between experiments.

The score T can be used to rank models. We prove mathematically that with an increasing number of realizations and a finer time resolution, truly invariant models will indeed receive a higher rank than noninvariant models (*Materials and Methods*).

Stable Variable Ranking Procedure. In biological applications, modeling kinetic systems is a common approach that is used to generate hypotheses related to causal relationships between specific variables, e.g., to find species involved in the regulation of a target protein. The noninvariance score can be used to construct a stability ranking of individual variables. The ranking we propose is similar to Bayesian model averaging (BMA) (31) and is based on how often each variable appears in a top-ranked model. The key advantage of such a ranking is that it leverages information from several fits, leading to an informative ranking. It also allows testing whether a specific variable is ranked significantly higher than would be expected from a random ranking (*Materials and Methods*). Moreover, we provide a theoretical guarantee under which the top-ranked variables are indeed contained in the true causal model (*Materials and Methods*).

We compare the performance of this ranking on a simulation study based on a biological ordinary differential equation system from the BioModels Database (32) which describes reactions in heated monosaccharide–casein systems (*SI Appendix*, section 4). (In fact, the example in Fig. 2 comes from this model, with Y and X^8 being the concentrations of Melanoidin and AMP, respectively.) We compare our method to dynamic Bayesian networks (33) based on conditional independence (DBN-CondInd), gradient matching (GM), and an integrated version thereof, which from now on we refer to as difference matching (DM); the last 2 methods both use ℓ_1 penalization for regularization (*SI Appendix*, section 4). Fig. 3*A* shows median receiver operator curves (ROCs) for recovering the correct causal parents based on 500 simulations for all 4 methods. CausalKinetiX has the fastest recovery rate and, in more than 50% of the cases, it is able to recover all causal parents without making any false discoveries (Fig. 3*C*). The recovery of the causal parents as a function of noise level is given in Fig. 3*B*. On the x axis, we plot the relative size of the noise, where a value of 1 implies that the size of the noise is on the same level as the target dynamic and the signal is very weak. For all noise levels, CausalKinetiX is better at recovering the correct model than all competing methods. More comparisons can be found in *SI Appendix*, section 4.

Numerical Stability, Scalability, and Misspecified Models. The method CausalKinetiX builds on standard statistical procedures, such as smoothing, quadratic programming, and regression. As opposed to standard nonlinear least squares, it does not make use of any numerical integration techniques. This avoids computational issues that arise when the dynamics result in stiff systems (34). For each model, the runtime is less than cubic in the sample size, which means that the key computational cost is the exhaustive model search. We propose to use a screening step to reduce the number of possible models (*SI Appendix*, section 3C), which allows applying the method to systems with hundreds of variables (e.g., *Experiment on Metabolic Network* below). Moreover, it does not require any assumptions on the dynamics of the covariates. In this sense, the method is robust with respect to model misspecifications on the covariates that can originate from hidden variables or misspecified functional relationships. Consistency of the proposed variable ranking (*SI Appendix*, section 3), for example, requires only the model for the target variable to be correctly specified. Simulation experiments show that this robustness can be observed empirically (*SI Appendix*, section 4). Finally, there is empirical evidence that incorporating invariance can be interpreted as regularization preventing overfitting and that the method is robust against correlated measurement error (*SI Appendix*, section 4).

Generalization in Metabolic Networks. We apply the proposed method to a real biological dataset of a metabolic network (*Materials and Methods*). Ion counts of one target variable and cell concentrations of 411 metabolites are measured at 11 time points across 5 different experimental conditions, each of which contains 3 biological replicates. The experiments include both up- and downshifts of the target variable; i.e., some of the conditions induce an increase of the target trajectory, compared to its starting value, and other conditions induce a decrease.

We compare the performance of CausalKinetiX with the performance of nonlinear least squares (NONLSQ). To make the methods feasible for such a large dataset, we combine them with a screening based on DM. We thus call the method based

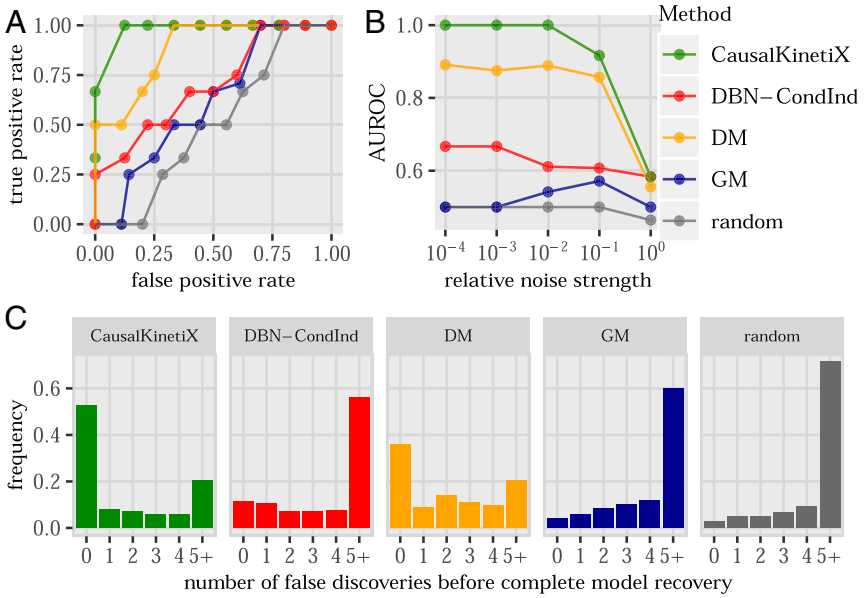

Fig. 3. (*A*) Median ROCs for recovering the correct causal parents based on 500 simulations. CausalKinetiX has the fastest recovery rate. (*B*) Median area under the receiver operator curve (AUROC) for different relative noise levels. CausalKinetiX outperforms all other methods. (*C*) Number of recoveries before all correct variables enter the model. In the majority of cases, CausalKinetiX has no false discovery.

on nonlinear least squares DM-NONLSQ; its parameters are estimated using the software Data2Dynamics (d2d) (14), which uses CVODES of the SUNDIALS suite (35) for numerical integration.

Fig. 4*A* shows the models' ability to describe the dynamics in the observed experiments (in-sample performance). DM-NONLSQ directly optimizes the RSS and therefore fits the data better than CausalKinetiX, which takes into account stability, as well. The RSS for DM-NONLSQ-10 (based on 10 terms) is lower (0.83) compared to CausalKinetiX (0.96) averaged over all in-sample experiments. The plot contains diagnostics for analyzing kinetic models. The integrated dynamics are shown jointly with a smoother (blue) through the observations (gray). At the observed time points, the predicted derivatives (red lines) are also shown using smoothed X and Y values. Model fits that explain the data reasonably well in the sense that the integrated trajectory is not far from the observations may predict derivatives (short red lines) on the smoother that do not agree well with the data: They fail to explain the underlying dynamics. For an example, see the in-sample fit of DM-NONLSQ-10 in Fig. 4*A*. We regard plotting a smoothing spline and the predicted derivatives for the fitted values as a highly informative tool when analyzing models for kinetic systems.

Pooling data across heterogeneous experiments, as, for example, done by DM-NONLSQ, is already a natural regularization technique; if there is sufficient heterogeneity in the data, the causal model is the only invariant model. Finitely many experiments, however, exhibit only limited heterogeneity and one can benefit from focusing specifically on invariant models. To compare the out-of-sample performance of the methods, we consider the best-ranked model from Fig. 4*A*, hold out 1 experiment, fit the parameters on the remaining 4 experiments, and predict the dynamics on the held-out experiment. While DM-NONLSQ-10 explains the observations well in-sample, it does not generalize to the held-out experiments, and neither does DM-NONLSQ-3 (based only on 3 terms), which avoids overfitting. The average RSS of the held-out experiments are 1.41, 2.95, and 3.45 for CausalKinetiX, DM-NONLSQ-10, and DM-NONLSQ-3, respectively (Fig. 4*B*). Another comparison, when the methods are fully agnostic about one of the experimental conditions, is provided in *SI Appendix*, section 4. By trading off invariance and predictability, CausalKinetiX yields models that perform well on unseen experiments that have not been used for parameter estimation.

Discussion

In the natural sciences, differential equation modeling is a widely used tool for describing kinetic systems. The discovery and verification of such models from data have become a fundamental challenge of science today. Existing methods are often based on standard model selection techniques or various types of sparsity enforcing regularization; they usually focus on predictive performance and sometimes consider stability with respect to resampling (36, 37). In this work, we develop methodology for structure search in ordinary differential equation models. Exploiting ideas from causal inference, we propose to rank models not only by their predictive performance, but also by taking into account invariance, i.e., their ability to predict well in different experimental settings. Based on this model ranking, we construct a ranking of individual variables reflecting causal importance. It provides researchers with a list of promising candidate variables that may be investigated further by performing interventional experiments, for example. Our ranking methodology (both for models and for variables) comes with theoretical asymptotic guarantees and with a clear statement of the required assumptions. Extensive experimental evaluation on simulated data shows that our method is able to outperform current state-of-the art methods. Practical applicability of the procedure is further illustrated on a not yet published biological dataset. Our implementation is readily available as an open-source R package (38).

The principle of searching for invariant models opens up a promising direction for learning causal structure from realistic, heterogeneous datasets. The proposed CausalKinetiX framework is flexible in that it can be combined with a wide range of dynamical models and any parameter inference method. This is particularly relevant when the differential equations depend nonlinearly on the parameters. Future extensions may further include the extension to stochastic, partial, and delay

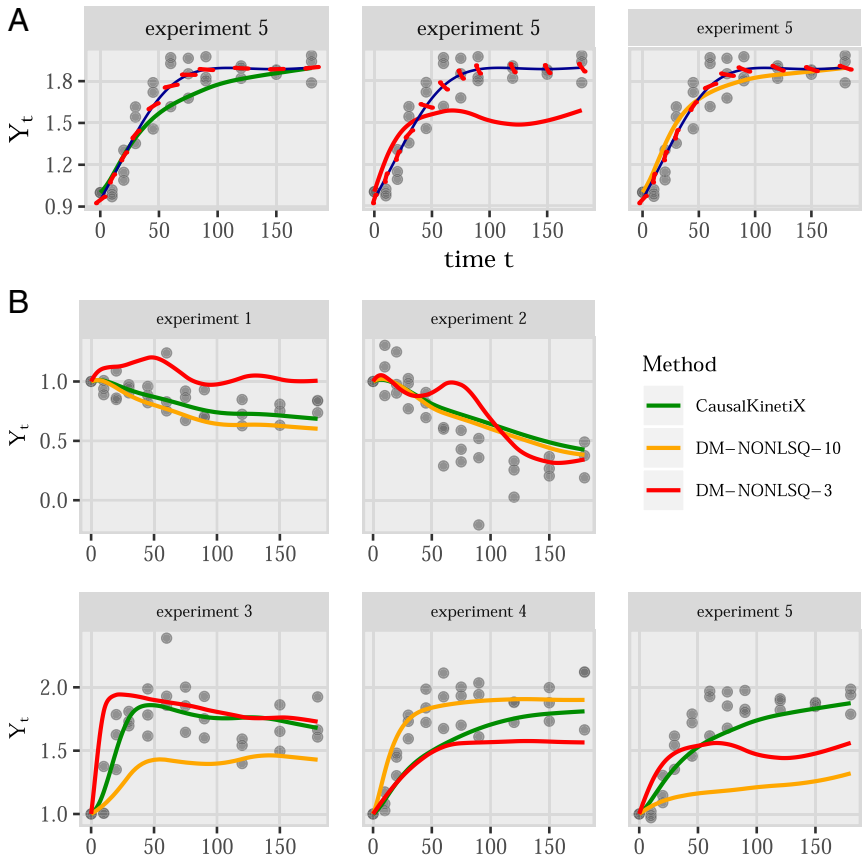

Fig. 4. Metabolic network analysis. (*A*) In-sample fit. All 5 experiments are used for model selection and parameter estimation. The plot shows model-based trajectories (numerically integrated) for experiment 5. DM with 10 terms (*Right*) fits the data better than CausalKinetiX (*Left*) or DM with 3 terms (*Center*). (*B*) Out-of-sample fit. The plot shows the models' ability to generalize to new experiments. Each plot shows model-based trajectories that are obtained when that experiment is not used for parameter estimation. CausalKinetiX shows the best generalization performance. The large DM model does not generalize well to unseen data due to overfitting.

differential equations and the transfer to other areas of application like robotics, climate sciences, neuroscience, and economics.

Materials and Methods

In this section, we provide additional details about data format, methodology, and the experiments. We further prove that our method is statistically consistent; i.e., it infers correct models when the sample size grows toward infinity.

Data Format. The data consist of n repetitions of discrete time observations of the d variables $\mathbf{X}$ (or their noisy version $\widetilde{\mathbf{X}}$) on the time grid $\mathbf{t} = (t_1, \dots, t_L)$. Each of the repetitions is part of an experiment $\{e_1, \dots, e_m\}$. The experiments should be thought of as different states of the system and may stem from different interventions. One of the variables, X^1, say, is considered as the target, and we write $Y_t^{1,(i)} = X_t^{1,i}$. We further assume an underlying dynamical model (which then results in various statistical dependencies between the variables and different time points).

Mass-Action Kinetic Models. Many ordinary differential equation-based systems in biology are described by the law of mass-action kinetics. The resulting ordinary differential equation models are linear combinations of various orders of interactions between the predictor variables $\mathbf{X}$. Assuming that the underlying ordinary differential equation model of our target $Y = X^1$ is described by a version of the mass-action kinetic law, the derivative $\dot{Y}_t := \frac{d}{dt} Y_t$ equals

$$\dot{Y}_t = g_\theta(\mathbf{X}_t) = \sum_{k=1}^{d} \theta_{0,k} X_t^k + \sum_{j=1}^{d} \sum_{k=j}^{d} \theta_{j,k} X_t^j X_t^k, \qquad [2]$$

where $\theta = (\theta_{0,1}, \dots, \theta_{0,d}, \theta_{1,1}, \theta_{1,2}, \dots, \theta_{d,d}) \in \mathbb{R}^{d(d+1)/2+d}$ is a parameter vector. We denote the subclass of all such linear models of degree 1 consisting of at most p terms (i.e., p nonzero terms in the parameter vector θ) by $\mathcal{M}_p^{\text{Exhaustive}}$ and call these models exhaustive linear models of degree 1. A more detailed overview of different collections of models is given in *SI Appendix*, section 2.

Model Scoring. For each model M the score $T = T(M)$ is computed using the following steps. They include fitting 2 models to the data: one in M3 and the other one in M4 and M5.

M1) Input: Data are as described above and a collection $\mathcal{M} = \{M^1, M^2, \dots, M^m\}$ of models over d variables is assumed to be rich enough to describe the desired kinetics. In the case of mass-action kinetics, e.g., $\mathcal{M} = \mathcal{M}_p^{\text{Exhaustive}}$.

M2) Screening of predictor terms (optional): For large systems, reduce the search space to fewer predictor terms. Essentially, any variable reduction technique based on the regression in step M4 can be used. We propose using ℓ_1-penalized regression (*SI Appendix*, section 2).

M3) Smooth target trajectories: For each repetition $i \in \{1, \dots, n\}$, smooth the (noisy) data $\widetilde{Y}_{t_1}^{(i)}, \dots, \widetilde{Y}_{t_L}^{(i)}$ using a smoothing spline

$$\hat{y}_a^{(i)} := \underset{y \in \mathcal{H}_C}{\operatorname{argmin}} \sum_{\ell=1}^{L} \left(\widetilde{Y}_{t_\ell}^{(i)} - y(t_\ell) \right)^2 + \lambda \int \ddot{y}(s)^2 \, ds, \qquad [3]$$

where λ is a regularization parameter, which in practice is chosen using cross-validation; $\mathcal{H}_C$ contains all smooth functions $[0, T] \to \mathbb{R}$, for which values and first 2 derivatives are bounded in absolute value by C. We denote the resulting functions by $\hat{y}_a^{(i)} : [0, T] \to \mathbb{R}$, $i \in \{1, \dots, n\}$. For each of the m candidate target models $M \in \mathcal{M}$ perform steps M4 to M6.

M4) Fit candidate target model: For every $i \in \{1, \dots, n\}$, find the function $g^i \in \mathcal{G}$ such that

$$\dot{Y}_t^{(k)} = g^i\left(\mathbf{X}_t^{(k)}\right) \quad [4]$$

is satisfied as well as possible for all $t \in \mathbf{t}$ and for all repetitions k belonging to a different experiment than repetition i. Below, we describe 2 procedures for this estimation step resulting in estimates $\hat{g}^i$. For each repetition $i \in \{1, \dots, n\}$, this yields L fitted values $\hat{g}^i(\widetilde{\mathbf{X}}_{t_1}^{(i)}), \dots, \hat{g}^i(\widetilde{\mathbf{X}}_{t_L}^{(i)})$. Leaving out the experiment of repetition i ensures that only an invariant model leads to a good fit, as these predicted derivatives are reasonable only if the dynamics generalize across experiments.

M5) Smooth target trajectories with derivative constraint: Refit the target trajectories for each repetition $i \in \{1, \dots, n\}$ by constraining the smoother to these derivatives; i.e., find the functions $\hat{y}_b^{(i)} : [0, T] \to \mathbb{R}$ which minimize

$$\hat{y}_b^{(i)} := \underset{y \in \mathcal{H}_C}{\arg\min} \sum_{\ell=1}^{L} \left(\widetilde{Y}_{t_\ell}^{(i)} - y(t_\ell)\right)^2 + \lambda \int \ddot{y}(s)^2 \, ds, \quad \text{such that} \quad \dot{y}(t_\ell) = \hat{g}^i(\widetilde{\mathbf{X}}_{t_\ell}^{(i)}) \text{ for all } \ell = 1, \dots, L. \quad [5]$$

M6) Compute score: If the candidate model M allows for an invariant fit, the fitted values $\hat{g}^i(\widetilde{\mathbf{X}}_1^{(i)}), \dots, \hat{g}^i(\widetilde{\mathbf{X}}_L^{(i)})$ computed in M4 will be reasonable estimates of the derivatives $\dot{Y}_{t_1}^{(i)}, \dots, \dot{Y}_{t_L}^{(i)}$. This, in particular, means that the constrained fit in M5 will be good, too. If, conversely, the candidate model M does not allow for an invariant fit, the estimates produced in M4 will be poor. We thus score the models by comparing the fitted trajectories $\hat{y}_a^{(i)}$ and $\hat{y}_b^{(i)}$ across repetitions as

$$T(M) := \frac{1}{n} \sum_{i=1}^{n} \left[\mathrm{RSS}_b^{(i)} - \mathrm{RSS}_a^{(i)}\right] / \left[\mathrm{RSS}_a^{(i)}\right], \quad [6]$$

where $\mathrm{RSS}_*^{(i)} := \frac{1}{L} \sum_{\ell=1}^{L} \left(\hat{y}_*^{(i)}(t_\ell) - \widetilde{Y}_{t_\ell}^{(i)}\right)^2$. If there is a reason to believe that the observational noise has similar variances across experiments, the division in the score can be removed to improve numerical stability.

The scores $T(M)$ induce a ranking on the models $M \in \mathcal{M}$, where models with a smaller score have more stable fits than models with larger scores. Below, we show consistency of the model ranking.

Variable Ranking. The following method ranks individual variables according to their importance in obtaining invariant models. We score all models in the collection $\mathcal{M}$ based on their stability score $T(M)$ (see **[6]**) and then rank the variables according to how many of the top-ranked models depend on them. This can be summarized in the following steps:

V1) Input: Same as in M1.

V2) Compute stabilities: For each model $M \in \mathcal{M}$ compute the noninvariance score $T(M)$ as described in **[6]**. Denote by $M_{(1)}, \dots, M_{(K)}$ the K top-ranked models, where $K \in \mathbb{N}$ is chosen to be the number of expected invariant models in $\mathcal{M}$.

V3) Score variables: For each variable $j \in \{1, \dots, d\}$, compute the following score:

$$s_j := \frac{|\{k \in \{1, \dots, K\} \mid M_{(k)} \text{ depends on } j\}|}{K}. \quad [7]$$

Here, "$M_{(k)}$ depends on j" means that the variable j has an effect in the model $M_{(k)}$ (*SI Appendix*, section 2). If there are K invariant models, the above score represents the fraction of invariant models that depend on variable j. It equals 1 for variable j if and only if every invariant model depends on that variable.

These scores s^j are similar to what is referred to as inclusion probabilities in Bayesian model averaging (31). Below, we construct hypothesis tests for the test whether a score is significantly higher than if the models are ranked randomly.

A natural choice for the parameter K should equal the number of invariant models. This may be unknown in practice, but our empirical studies found that the method's results are robust to the choice of K. In particular, we propose to choose a small K to ensure that it is smaller than the number of invariant models (*SI Appendix*, section 2).

Fitting Target Models (M4). In step M4, for every $i \in \{1, \dots, n\}$, we perform a regression to find a function $g^i \in M$ such that **[4]** is optimized across all repetitions k belonging to different experiments than i. This task is difficult for 2 reasons. First, the derivative values $\dot{Y}_t^{(k)}$ are not directly observed and, second, even if we had access to (noisy and unbiased versions of) $\dot{Y}_t^{(k)}$, we are dealing with an error-in-variables problem. Nevertheless, for certain model classes it is possible to perform this estimation consistently and since the predictions are used only as constraints, one expects estimates to work as long as they preserve the general dynamics. We propose 2 procedures: 1) a general method that can be adapted to many model classes and 2) a method that performs better but assumes the target model to be linear in parameters.

The first procedure estimates the derivatives and then performs a regression based on the model class under consideration. That is, one fits the smoother $y_a^{(k)}$ from M3 and then computes its derivatives. When using the first derivative of a smoothing spline, it has been argued that the penalty term in **[3]** contains the third rather than the second derivative of y (39). We then regress the estimated derivatives on the data. As a regression procedure, one can use ordinary least squares if the models are linear or random forests, for example, if the functions are highly nonlinear.

The second method works for models that are linear in the parameters, i.e., for models that consist of functions of the form $g_\theta(\mathbf{x}) = \sum_{j=1}^{p} \theta_j g_j(\mathbf{x})$, where the functions $g_1, \dots, g_p$ are known transformations. This yields

$$Y_{t_\ell}^{(k)} - Y_{t_{\ell-1}}^{(k)} = \sum_{j=1}^{p} \theta_j \int_{t_{\ell-1}}^{t_\ell} g_j(\widetilde{\mathbf{X}}_s^{(k)}) ds.$$

This approach does not require estimation of the derivatives of Y but instead uses the integral of the predictors. It is well known that integration is numerically more stable than differentiation (40). Often, it suffices to approximate the integrals using the trapezoidal rule; i.e.,

$$\int_{t_{\ell-1}}^{t_\ell} g_j(\widetilde{\mathbf{X}}_s^{(k)}) ds \approx \frac{g_j(\widetilde{\mathbf{X}}_{t_\ell}^{(k)}) + g_j(\widetilde{\mathbf{X}}_{t_{\ell-1}}^{(k)})}{2} (t_\ell - t_{\ell-1}),$$

since the noise in the predictors is often stronger than the error in this approximation. The resulting bias is then negligible.

As mentioned above, most regression procedures have difficulties with errors in variables and therefore return biased results. Sometimes it can therefore be helpful to use smoothing or averaging of the predictors to reduce the impact of this problem. Our procedure is flexible in the sense that other fitting procedures, e.g., inspired by refs. 28, 41, and 42, could be applied, too.

Experiment on Metabolic Network. Defining the auxiliary variable $Z_t := 2 - Y_t$, we expect that the target species Y_t and Z_t are tightly related: $Y_t \rightleftharpoons Z_t$, i.e., Y_t is formed into Z_t and vice versa. We therefore expect models of the form

$$\dot{Y}_t = \theta_1 Z_t X_t^j X_t^k + \theta_2 Z_t X_t^p X_t^q - \theta_3 Y_t X_t^r X_t^s$$
$$\dot{Z}_t = -\theta_1 Z_t X_t^j X_t^k - \theta_2 Z_t X_t^p X_t^q + \theta_3 Y_t X_t^r X_t^s,$$

where $j, k, p, q, r, s \in \{1, \dots, 411\}$ and $\theta_1, \theta_2, \theta_3 \geq 0$. By the conservation of mass both target equations mirror themselves, which makes it sufficient to learn only the model for Y_t. More precisely, we use the model class consisting of 3-term models of the form $Z_t X_t^j X_t^k$, $Y_t X_t^j X_t^k$, $Z_t X_t^j$, $Y_t X_t^j$, Z_t, or Y_t, where the sign of the parameter is constrained to being positive or negative depending on whether the term contains Z_t or Y_t, respectively. We constrain ourselves to 3 terms, as we found this to be the smallest number of terms that results in sufficiently good in-sample fits. Given sufficient computational resources, one may include more terms, too, of course. The sign constraint can be incorporated into our method by performing a constrained least-squares fit instead of OLS in step M4. This constrained regression can then be solved efficiently by a quadratic program with linear constraints.

As the biological data are high dimensional, our method first screens down to 100 terms and then searches over all models consisting of 3 terms. To get more accurate fits of the dynamics, we pool and smooth over the 3 biological replicates and work only with the smoothed data.

Significance of Variable Ranking. We can test whether a given score s_j, defined in **[7]**, is significant in the sense that the number of top-ranked models depending on variable j is higher than one would expect if the

ranking of all models in $\mathcal{M}$ was random. More precisely, consider the null hypothesis

$$H_0: \quad \begin{array}{c}\text{the top-ranked models } M_{(1)}, \ldots, M_{(K)} \\ \text{are drawn uniformly from all models in } \mathcal{M}.\end{array}$$

It is straightforward to show that under H_0 it holds that $K \cdot s_j$ follows a hypergeometric distribution with parameters $|\mathcal{M}|$ (population size), $|\{M \in \mathcal{M} \mid M \text{ depends on } j\}|$ (number success in population), and K (number of draws). For each variable we can hence compute a p value to assess whether it is significantly important for stability.

Theoretical Consistency Guarantees. We prove that both the model ranking and the proposed variable ranking satisfy theoretical consistency guarantees. More precisely, under suitable conditions and in the asymptotic setting where both the number of realizations n and the number of time points L converge to infinity, every invariant model will be ranked higher than all noninvariant models. Given sufficient heterogeneity of the experiments it additionally holds that the variable score s_j defined in **[7]** tends to one if and only if $j \in S^*$ (see **[1]**). Details and proofs are provided in *SI Appendix*, section 3.

Relation to Causality. Causal models enable us to model a system's behavior not only in an observational state, but also under interventions. There are various ways to define causal models (22, 23). The concept of structural causal models is well suited for the setting of this paper and its formalism can be adapted to the case of dynamical models (*SI Appendix*, section 1). If the experimental settings correspond to different interventions on variables other than Y, choosing S^* as the set of causal parents of Y satisfies **[1]**. If the settings are sufficiently informative, no other set satisfies **[1]**.

Code and Data Availability. Well-documented code is available as an open-source R package on CRAN (38). It includes the ordinary differential equation models used in the simulations, e.g., the Maillard reaction. All further code and data are available at http://CausalKinetiX.org. The exact parameter settings for the simulations can be found in *SI Appendix*, section 4. The data underlying the metabolic network experiment will be made available upon reasonable request.

ACKNOWLEDGMENTS. We thank R. Loewith, B. Ryback, U. Sauer, E. M. Sayas, and J. Stelling for providing the biological dataset as well as helpful biological insights. We further thank N. R. Hansen and N. Meinshausen for helpful discussions; and K. Ishikawa and A. Orvieto for their help with Python and d2d.

1. K. Friston, L. Harrison, W. Penny, Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
2. T. Chen, H. He, G. Church, "Modeling gene expression with differential equations" in *Biocomputing'99*, R. B. Altman, K. Lauderdale, A. K. Dunker, L. Hunter, T. E. Klein, Eds. (World Scientific, 1999), pp. 29–40.
3. B. Ogunnaike, W. Ray, *Process Dynamics, Modeling, and Control* (Oxford University Press New York, NY, 1994), vol. 1.
4. R. Murray, *A Mathematical Introduction to Robotic Manipulation* (CRC Press, 2017).
5. W.-B. Zhang, *Differential Equations, Bifurcations, and Chaos in Economics* (World Scientific Publishing Company, 2005), vol. 68.
6. D. Fey *et al.*, Signaling pathway models as biomarkers: Patient-specific simulations of jnk activity predict the survival of neuroblastoma patients. *Sci. Signal.* **8**, ra130 (2015).
7. H. Hass *et al.*, Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ Syst. Biol. Appl.* **3**, 27 (2017).
8. C. L. Arteaga, J. A. Engelman, Erbb receptors: From oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer Cell* **25**, 282–303 (2014).
9. S.-X. Ren *et al.*, Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature* **422**, 888–893 (2003).
10. A. Regev *et al.*, Science forum: The human cell atlas. *Elife* **6**, e27041 (2017).
11. J. Rozman *et al.*, Identification of genetic elements in metabolism by high-throughput mouse phenotyping. *Nat. Commun.* **9**, 288 (2018).
12. Y. Bard, *Nonlinear Parameter Estimation* (Academic Press, New York, NY, 1974).
13. M. Benson, Parameter fitting in dynamic models. *Ecol. Model.* **6**, 97–115 (1979).
14. A. Raue *et al.*, Data2dynamics: A modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics* **31**, 3558–3560 (2015).
15. B. Schölkopf *et al.*, "On causal and anticausal learning" in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, J. Langford, J. Pineau, Eds. (Omnipress, Madison, WI, 2012), pp. 459–466.
16. B. Yu, Stability. *Bernoulli* **19**, 1484–1500 (2013).
17. J. Peters, P. Bühlmann, N. Meinshausen, Causal inference using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B* **78**, 947–1012 (2016).
18. E. Bareinboim, J. Pearl, Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7345–7352 (2016).
19. N. Meinshausen *et al.*, Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7361–7368 (2016).
20. R. M. Shiffrin, Drawing causal inference from big data. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7308–7309 (2016).
21. B. Yu, K. Kumbier, Veridical Data Science (PCS). arXiv:1901.08152 (12 November 2019).
22. J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, New York, NY, ed. 2, 2009).
23. G. W. Imbens, D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, New York, NY, 2015).
24. T. Haavelmo, The probability approach in econometrics. *Econometrica* **12**(suppl.), S1–S115 (1944).
25. J. Aldrich, Autonomy. *Oxf. Econ. Pap.* **41**, 15–34 (1989).
26. J. Pearl, D. Mackenzie, *The Book of Why* (Basic Books, New York, NY, 2018).
27. J. Peters, D. Janzing, B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms* (MIT Press, Cambridge, MA, 2017).
28. C. J. Oates *et al.*, Causal network inference using biochemical kinetics. *Bioinformatics* **30**, i468–i474 (2014).
29. D. Eaton, K. P. Murphy, "Exact Bayesian structure learning from uncertain interventions" in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, M. Meila, X. Shen, Eds. (Journal of Machine Learning Reseach [JMLR], 2007), pp. 107–114.
30. C. Lim, B. Yu, Estimation stability with cross-validation (ESCV). *J. Comput. Graph. Stat.* **25**, 464–492 (2016).
31. J. Hoeting, D. Madigan, A. Raftery, C. Volinsky, Bayesian model averaging: A tutorial. *Stat. Sci.* **14**, 382–417 (1999).
32. C. Li *et al.*, Biomodels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* **4**, 92 (2010).
33. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
34. L. F. Shampine, *Numerical Solution of Ordinary Differential Equations* (Routledge, 2018).
35. A. C. Hindmarsh *et al.*, SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math Software* **31**, 363–396 (2005).
36. N. Meinshausen, P. Bühlmann, Stability selection. *J. R. Stat. Soc. Ser. B* **72**, 417–473 (2010).
37. S. Basu, K. Kumbier, J. B. Brown, B. Yu, Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 1943–1948 (2018).
38. N. Pfister, S. Bauer, J. Peters, CausalKinetiX: Learning Stable Structures in Kinetic Systems. CausalKinetiX. https://cran.r-project.org/web/packages/CausalKinetiX. Deposited 20 June 2019.
39. J. O. Ramsay, B. W. Silverman, *Functional Data Analysis* (Springer, New York, NY, 2005).
40. S. Chen, A. Shojaie, D. M. Witten, Network reconstruction from high-dimensional ordinary differential equations. *J. Am. Stat. Assoc.* **112**, 1697–1707 (2017).
41. J. O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B* **69**, 741–796 (2007).
42. B. Calderhead, M. Girolami, N. D. Lawrence, "Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes" in *Advances in Neural Information Processing Systems (NIPS)*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta, Eds. (Curan, 2009), pp. 217–224.